What is Privacy?

Webster:  

*Freedom from unauthorized intrusion*

- Intrusive
  - Is disclosure of the data not in the individual’s best interest?
Intrusion

• Harm to individual
  – Physical, psychological, or perceived
  – How to measure?
• Use of data for other than approved purpose
  – Current standard in many areas
  – Too restrictive?
  – Too lenient?

Privacy

• “the ability to access and control one's personal information”
• Recognized by several treaties and protected by law
  – United States Healthcare Insurance Portability and Accountability (HIPAA)
  – The European Community Directive 95/46/EC
  – Privacy is about “individually identifiable data”
Terminology

- **Private Data**
  - Individually Identifiable
  - Sensitive
- **Parties**
  - Data subject
    - Person who the private data is about
  - Processor
    - Handles/manages private data
  - Recipient
    - Someone to whom data is disclosed
  - Adversary
    - One who would/could misuse private data

Regulatory Constraints: Privacy Rules

- Primarily national laws
  - European Union
  - US HIPAA rules ([www.hipaadvisory.com](http://www.hipaadvisory.com))
  - Many others: ([www.privacyexchange.org](http://www.privacyexchange.org))
- Often control transborder use of data
- Focus on intent
  - Limited guidance on implementation
European Union Data Protection Directives

- Directive 95/46/EC
  - Passed European Parliament 24 October 1995
  - Goal is to ensure free flow of information
    - Must preserve privacy needs of member states
  - Effective October 1998
- Effect
  - Provides guidelines for member state legislation
    - Not directly enforceable
  - Forbids sharing data with states that don’t protect privacy
    - Non-member state must provide adequate protection,
    - Sharing must be for “allowed use”, or
    - Contracts ensure adequate protection
  - US “Safe Harbor” rules provide means of sharing (July 2000)
    - Adequate protection
    - But voluntary compliance
- Enforcement is happening
  - Microsoft under investigation for Passport (May 2002)
  - Already fined by Spanish Authorities (2001)

EU 95/46/EC: Meeting the Rules

- Personal data is any information that can be traced directly or indirectly to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn’t violate privacy
EU 95/46/EC: Meeting the Rules

- Some uses specifically proscribed
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life
- Must make data available to subject
  - Allowed to object to such use
  - Must give advance notice / right to refuse direct marketing use
- Limits use for automated decisions (e.g., creditworthiness)
  - Person can opt-out of automated decision making
  - Onus on processor to show use is legitimate and safeguards in place to protect person’s interests
  - Logic involved in decisions must be available to affected person
- [europa.eu.int/comm/internal_market/privacy/index_en.htm](http://europa.eu.int/comm/internal_market/privacy/index_en.htm)

US Health Insurance Portability and Accountability Act (HIPAA)

- Governs use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual
    - For treatment (generally requires consent)
    - To public health / legal authorities
  - Use permitted where “there is no reasonable basis to believe that the information can be used to identify an individual”
- Safe Harbor Rules
  - Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
    - Name, location smaller than 3 digit postal code, dates finer than year, identifying numbers
  - Shown not to be sufficient (Sweeney)
  - Also not necessary
  - Moral: Get Involved in the Regulatory Process!
Contractual Limitations

- Web site privacy policies
  - “Contract” between browser and web site
  - Groups support voluntary enforcement
    - TrustE – requires that web site DISCLOSE policy on collection and use of personal information
    - BBBonline
      - posting of an online privacy notice meeting rigorous privacy principles
      - completion of a comprehensive privacy assessment
      - monitoring and review by a trusted organization, and
      - participation in the programs consumer dispute resolution system
    - Unknown legal “teeth”
      - Example of customer information viewed as salable property in court!!!
  - P3P: Supports browser checking of user-specific requirements
    - Internet Explorer 6 – disallow cookies if non-matching privacy policy
    - PrivacyBird – Internet Explorer plug-in from AT&T Research

- Corporate agreements
  - Stronger teeth/enforceability
  - But rarely protect the individual

Defining Privacy
Modeling Real World

- What type of data the owner has?
  - Single table, relational, spatio-temporal, transactional, stream…

- What does the adversary know?
  - External public tables, phone books, names, ages, addresses…

- What is sensitive?
  - Medical history, salary, GPA…

- What is the RISK OF DISCLOSURE on both subject’s end and owner’s end?
  - Discrimination, public humiliation…
  - Court suits
Anonymization

- Goal: Not individually identifiable data
  - Specifically exempt from privacy laws
- Approaches
  - Remove identifiers
  - Generalization/suppression of non-identifiers
- Sensitive values still correct/usable
  - But what if generalized/suppressed values needed?

A Bogus Real World Model

- Data owner, hospital, has medical records
- Adversary knows names of the subjects
- Disease information is sensitive

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obi</td>
<td>17</td>
<td>M</td>
<td>Turkey</td>
<td>Flu</td>
</tr>
<tr>
<td>Leia</td>
<td>16</td>
<td>F</td>
<td>Bulgaria</td>
<td>Flu</td>
</tr>
<tr>
<td>Padme</td>
<td>23</td>
<td>F</td>
<td>US</td>
<td>Obesity</td>
</tr>
<tr>
<td>Yoda</td>
<td>25</td>
<td>M</td>
<td>Canada</td>
<td>Tetanus</td>
</tr>
</tbody>
</table>

Solution: Remove Unique Identifiers
Model Fails

Quasi Identifiers

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>M</td>
<td>Turkey</td>
<td>Flu</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>Bulgaria</td>
<td>Flu</td>
</tr>
<tr>
<td>23</td>
<td>F</td>
<td>US</td>
<td>Obesity</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>Canada</td>
<td>Tetanus</td>
</tr>
</tbody>
</table>

Private Dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Nation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obi</td>
<td>17</td>
<td>M</td>
<td>Turkey</td>
</tr>
<tr>
<td>Leia</td>
<td>16</td>
<td>F</td>
<td>Bulgaria</td>
</tr>
<tr>
<td>Padme</td>
<td>23</td>
<td>F</td>
<td>US</td>
</tr>
<tr>
<td>Yoda</td>
<td>25</td>
<td>M</td>
<td>Canada</td>
</tr>
</tbody>
</table>

In the real world, an adversary might have access to unique and quasi identifiers of the subjects.

In US, postal code, gender, birth date unique for 87%

Re-identifying “anonymous” data (Sweeney ’01)

- 37 US states mandate collection of information
- She purchased the voter registration list for Cambridge Massachusetts – 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three

Solution: k-anonymity
  - Any combination of values appears at least k times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results